Dridi *et al. EURASIP Journal on Bioinformatics and Systems Biology* (2017) 2017:9
DOI 10.1186/s13637-017-0062-4

EURASIP Journal on Bioinformatics and Systems Biology

RESEARCH Open Access

Bayesian inference for biomarker discovery in proteomics: an analytic solution

Noura Dridi[1,2†], Audrey Giremus[1†], Jean-Francois Giovannelli[1*†], Caroline Truntzer[3], Melita Hadzagic[1,4], Jean-Philippe Charrier[5], Laurent Gerfault[6,7], Patrick Ducoroy[3], Bruno Lacroix[5], Pierre Grangeat[6,7] and Pascal Roy[8,9,10,11]

Abstract

This paper addresses the question of biomarker discovery in proteomics. Given clinical data regarding a list of proteins for a set of individuals, the tackled problem is to extract a short subset of proteins the concentrations of which are an indicator of the biological status (healthy or pathological). In this paper, it is formulated as a specific instance of variable selection. The originality is that the proteins are not investigated one after the other but the best partition between discriminant and non-discriminant proteins is directly sought. In this way, correlations between the proteins are intrinsically taken into account in the decision. The developed strategy is derived in a Bayesian setting, and the decision is optimal in the sense that it minimizes a global mean error. It is finally based on the posterior probabilities of the partitions. The main difficulty is to calculate these probabilities since they are based on the so-called evidence that require marginalization of all the unknown model parameters. Two models are presented that relate the status to the protein concentrations, depending whether the latter are biomarkers or not. The first model accounts for biological variabilities by assuming that the concentrations are Gaussian distributed with a mean and a covariance matrix that depend on the status only for the biomarkers. The second one is an extension that also takes into account the technical variabilities that may significantly impact the observed concentrations. The main contributions of the paper are: (1) a new Bayesian formulation of the biomarker selection problem, (2) the closed-form expression of the posterior probabilities in the noiseless case, and (3) a suitable approximated solution in the noisy case. The methods are numerically assessed and compared to the state-of-the-art methods (*t* test, LASSO, Battacharyya distance, FOHSIC) on synthetic and real data from proteins quantified in human serum by mass spectrometry in selected reaction monitoring mode.

Keywords: Variable selection, Model selection, Optimal decision, Bayesian approach, Evidence, Hierarchical model, Proteomics, Biomarker

1 Introduction

It is now generally recognized that protein expression analysis is crucial in explaining the changes that occur as a part of disease pathogenesis [1, 2]. In this context, recent advances in mass spectrometry (MS) technologies have facilitated the investigation of proteins over a wide range of molecular weights in small biological specimens from blood or urine samples, for instance. Notably, MS in selected reaction monitoring (SRM) mode has demonstrated its ability to quantify clinical biomarkers in patient sera [3, 4]. Consequently, a large amount of research has been generated in proteomics based on data such as protein mass spectral intensities or protein concentrations obtained from the spectra. Specifically, the focus is on the selection (or discovery) of the "signature profiles," the so-called biomarkers. They represent, for instance, indicators of normal versus pathogenic biological processes, or positive versus negative pharmacological responses to therapeutic intervention.

Critical to the identification of biomarkers are: (1) the biological variability, i.e., the random variations of the concentrations of proteins between individuals sharing

*Correspondence: Giova@IMS-Bordeaux.fr
†Equal contributors
[1]IMS (Univ. Bordeaux, CNRS, BINP), 33400 Talence, France
Full list of author information is available at the end of the article

the same biological status [5], and (2) the technical variability, which originates from the imperfections of the measurement process used to obtain the concentrations. Failing to address both of these variabilities within a technique for biomarker identification may significantly impair its performance by resulting in erroneous decision.

Furthermore, since the complexity of a status is unlikely to be manifested through the changes in the characteristics of just one protein, it has generally been acknowledged that a set of proteins should be considered [5–8]. An additional difficulty is that they are possibly correlated, imposing the use of multivariate models to account for all the data simultaneously. These aforementioned issues pose significant challenges in developing efficient and robust statistical techniques for the identification of biomarkers.

The paper tackles the problem of biomarker identification by adopting a Bayesian approach to propose the selection of the optimal set of variables. By providing an elegant and mathematically rigorous framework for incorporating the data and the prior information within a joint probabilistic model, the Bayesian setting allows straightforward modeling of both the technical and the biological variabilities of the data.

The remainder of the paper is organized as follows. Section 2 summarizes the state-of-the-art variable selection methods, discusses their main challenges, and outlines our principal contributions. Section 3 presents the proposed formulation within the Bayesian framework, the proposed models for the data, and the decision strategy. Section 4 describes the data used in the numerical evaluations, together with the results and their analysis. Finally, conclusions are drawn in Section 5. A detailed description of the model and the derivation of the analytic solution is provided in Appendix.

2 Related work

The identification of biomarkers for diagnosis or prognosis can be classically formulated as a variable selection problem, and this problem has been paid a lot of attention as a specific instance of model choice. Various methodologies exist that can be broadly classified in two categories: the frequentist hypothesis testing and the Bayesian decision-making.

Frequentist hypothesis testing consists in deciding between two statements, classically referred to as the null and the alternative hypotheses, by comparing a function of the observed data to a threshold. The reader is invited to consult [9] for a comprehensive overview. Two closely related methods have been proposed. On the one hand, Neyman-Pearson tests are designed to ensure the so-called type I error. On the other hand, p values focus on how strongly the data reject the null hypothesis H_0 by evaluating the probability of obtaining a value as extreme as the observed one given H_0 is true. In biomarker discovery, a popular approach consists in testing a mean difference between the case and the negative controlled populations using the classical Students' t test or its variants [7]. The latter is a statistical hypothesis test which indicates whether the difference between two group means most likely reflects that they are samples of two different populations or, on the contrary, is merely explained by the sampling fluctuation. However, as the number of candidate biomarkers increases, multiple hypothesis testing is required resulting in a higher computational cost which may become prohibitive [10]. A first solution is to perform univariate tests for each protein.

This procedure requires an adapted control of the rate of type I errors in this particular setting where multiple hypothesis tests are conducted simultaneously. Two types of procedures were proposed for this purpose, namely the so-called family wise error rate or the false discovery rate [11, 12]. A common criticism of frequentist approaches is that they fail to take into account prior information about the problem at hand such as interdependencies between the different variables.

For a large number of candidate variables, regression analysis [13] provides an alternative to the abovementioned methods. The principle is to assume that a given outcome is related to a linear combination of a set of explanatory variables called the predictors. In proteomics, logistic regression models are considered that express the probability to have a disease as a function of the protein abundances [14–16]. Then, variable selection is classically performed using stepwise procedures that consists in successively adding or removing predictors, estimating the regression coefficients, and evaluating the goodness of fit of the subsequent model. Different criteria can be considered such as the R-squared, the adjusted R-squared, or the Akaike Information Criterion [17, 18]. Such techniques are referred to as backward elimination and forward selection, respectively [13]. However, these selection procedures are prone to overfitting and the variance of the parameter estimates becomes high in the presence of correlated predictors. Regularization methods alleviate these difficulties by considering the minimum of a penalized least squares error as estimate. Since the Ridge regression in 1970 [19], several algorithms have been proposed that differ between one another with respect to the considered penalization of the regression parameters. The well-known LASSO [20] considers a L_1-norm penalty and has the advantage of directly removing irrelevant predictors by shrinking their coefficients to zero. More recently, the elastic net [21] which combines the advantages of the Ridge and LASSO regressions has been proposed. In the presence of correlated variables, it outperforms LASSO by favoring the selection of sets of variables. A comparison between these methods in application to

genome selection is presented in [22]. Although widely used, regression analysis is based on an ad hoc model that may not reflect the physical nature of the observed data. Further, it does not explicitly accommodate correlations between the candidate biomarkers as well as measurement errors.

The Bayesian framework offers an alternative formulation of model selection. The candidate models are assigned prior probabilities that are combined with the likelihood function to yield the so-called posterior probability. The latter summarizes all the available information to make the decision. In this context, deciding in favor of the a posteriori most probable model is optimal in the sense that it minimizes the risk associated to the 0/1 cost-function. There have been a lot of debates over the use of Bayesian techniques in place of frequentist approaches, but they do not address exactly the same question. Frequentist methods are designed to test the departure of the data from a pre-defined null hypothesis. In contrast, Bayesian selection procedures evaluate the plausibility of a given hypothesis given a set of candidate hypotheses hence are conveniently well-suited to multiple hypotheses testing. Thus, non-nested models can be compared in a straightforward manner. Another fundamental difference is the treatment of unknown model parameters. In the frequentist approach, they are classically replaced by estimates whereas in the Bayesian formulation, they are integrated. The latter procedure has the advantage of automatically penalizing complex models, as discussed in [23], but often leads to intractable calculations. An additional advantage is that correlations between the model variables can be easily accounted for in the design of the prior distributions. As for the integration over the unknown parameters, several solutions have been developed. The Laplace approximation of the integrand leads to the well-known Bayesian information criterion (BIC). As an alternative, numerical integrations can be performed based on stochastic sampling techniques such as Markov Chain Monte Carlo (MCMC) methods [24]. Either across or within model-based techniques can be considered. In the first case, the model index is sampled jointly with the parameters conditionally upon the observations. A well-known algorithm is the Reversible-Jump MCMC but moves between the different parameter spaces are difficult to design. In the second case, posterior samples of the parameters are generated conditionally upon each candidate model and then used to evaluate the integrated likelihood, also called evidence [25]. Nevertheless, the harmonic mean-based estimator exhibits instabilities [26]. Applications of the MCMC Bayesian model selection methods in genomics can be found in [27, 28].

In this paper, a Bayesian setting is adopted to identify a set of protein biomarkers from experimental data consisting of measured protein concentrations and the associated biological statuses of a population of individuals. The novelty is that the decision is not made protein by protein. As an alternative, the problem is formulated as directly finding the best partition of the list of proteins into two subsets, namely discriminant and non-discriminant, in the sense that it yields the highest posterior probability. Regardless of their discriminative power, the proteins are assumed Gaussian distributed. However, for the subset of biomarkers, the parameters of the Gaussian distribution take different values depending on the biological status whereas this is not the case for the second subset of proteins. The preliminary version of this hierarchical model has been presented in [29]. Its advantages are threefold. First, it is not based on an ad hoc explanatory model unlike regression analysis. Second, the proteins within a given group are assumed a priori correlated and the dependency structure is integrated out along with the remainder of the unknown model parameters so that only the protein partition is estimated. Thus, our approach intrinsically takes into account correlations between the candidate biomarkers. Third, by choosing appropriate conjugate prior distributions for the parameters, the model evidences can be calculated in closed form and there is no need to resort to computationally extensive numerical techniques. Finally, we show that our hierarchical model can be easily extended to address errors in the measured concentrations.

3 Problem formulation, proposed models, and methods

To formulate the biomarker selection problem and construct its solution in the proposed framework, we first introduce the basic modeling for the relevant quantities/variables at hand: the biological status, protein concentrations, number of individuals,... including the descriptions of the considered observation models.

Distribution for status and concentration Regarding the biological status, it is denoted by b and takes two values, $\mathcal{H}$ and $\mathcal{P}$, for healthy and pathological. It is conveniently described by a Bernoulli random variable B with parameter p

$$B|p \sim \mathcal{B}(b\,;\,p)\,. \tag{1}$$

Regarding the proteins, let us note P is their number and $\boldsymbol{x} \in \mathbb{R}^P$ is the collection of their concentrations. Each protein can be discriminant or non-discriminant and then accordingly labeled by $+$ or $-$. The vector $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$, with sizes P^+ and P^- (we have $P = P^+ + P^-$), stand for the respective concentrations. Otherwise, within the P proteins, there are 2^P possible partitions referred to as $\boldsymbol{\delta} \in \{+,-\}^P$ since each protein can be discriminant and non-discriminant. Following the clinically observed

behavior, from a probabilistic standpoint, the concentrations are described by normal distributions in order to account for biological variabilities within the populations. Specifically, for the non-discriminant proteins, the concentration vector $\boldsymbol{x}^-$ is modeled by a unique multivariate normal distribution with common parameters $(\boldsymbol{m}_\mathcal{C}, \boldsymbol{\Gamma}_\mathcal{C})$ regardless the biological status:

$$\boldsymbol{X}^-|\boldsymbol{m}_\mathcal{C}, \boldsymbol{\Gamma}_\mathcal{C} \sim \mathcal{N}(\boldsymbol{x}^-; \boldsymbol{m}_\mathcal{C}, \boldsymbol{\Gamma}_\mathcal{C})\,. \tag{2}$$

On the other hand, for discriminant ones, the concentration vector $\boldsymbol{x}^+$ is modeled, conditionally on status b, by a multivariate normal distribution with mean and precision $(\boldsymbol{m}_\mathcal{H}, \boldsymbol{\Gamma}_\mathcal{H})$ and $(\boldsymbol{m}_\mathcal{P}, \boldsymbol{\Gamma}_\mathcal{P})$ for healthy and pathological, respectively:

$$\begin{cases} \boldsymbol{X}^+|b = \mathcal{P}, \boldsymbol{\delta}, \boldsymbol{m}_\mathcal{P}, \boldsymbol{\Gamma}_\mathcal{P} & \sim \mathcal{N}\left(\boldsymbol{x}^+;\, \boldsymbol{m}_\mathcal{P}, \boldsymbol{\Gamma}_\mathcal{P}\right), \\ \boldsymbol{X}^+|b = \mathcal{H}, \boldsymbol{\delta}, \boldsymbol{m}_\mathcal{H}, \boldsymbol{\Gamma}_\mathcal{H} & \sim \mathcal{N}\left(\boldsymbol{x}^+;\, \boldsymbol{m}_\mathcal{H}, \boldsymbol{\Gamma}_\mathcal{H}\right). \end{cases} \tag{3}$$

Marginally, the concentrations of discriminant proteins are distributed according to a mixture of two Gaussian distributions

$$\begin{aligned} \boldsymbol{X}^+|p, \boldsymbol{m}_\mathcal{P}, \boldsymbol{\Gamma}_\mathcal{P}, \boldsymbol{m}_\mathcal{H}, \boldsymbol{\Gamma}_\mathcal{H} &\sim p\,\mathcal{N}\left(\boldsymbol{x}^+; \boldsymbol{m}_\mathcal{H}, \boldsymbol{\Gamma}_\mathcal{H}\right) \\ &\quad + (1-p)\mathcal{N}\left(\boldsymbol{x}^+; \boldsymbol{m}_\mathcal{P}, \boldsymbol{\Gamma}_\mathcal{P}\right). \end{aligned}$$

In addition, it is assumed that $\boldsymbol{X}^+$ and $\boldsymbol{X}^-$ are conditionally independent.

The parameters of the distributions are collected in the vector $\boldsymbol{\theta} = \left[\boldsymbol{m}_\mathcal{P}, \boldsymbol{\Gamma}_\mathcal{P}, \boldsymbol{m}_\mathcal{H}, \boldsymbol{\Gamma}_\mathcal{H}, \boldsymbol{m}_\mathcal{C}, \boldsymbol{\Gamma}_\mathcal{C}, p\right]$ considered as unknown. It is important to keep in mind that the quantity of interest is the partition $\boldsymbol{\delta}$.

Distribution for the individuals The total number of individual is N and $(\boldsymbol{x}_n, b_n)$ is the nth concentration vector and status. They are modeled as independent conditionally on $\boldsymbol{\theta}$. Let us denote $\underline{\boldsymbol{x}}$ (size $P \times N$) as the matrix of concentrations and $\boldsymbol{b}$ (size N) as the vector of biological statuses. Also, let $\mathcal{I}_\mathcal{H}$ and $\mathcal{I}_\mathcal{P}$ be the subsets of indices for healthy and pathological individuals, respectively, and $N_\mathcal{H}, N_\mathcal{P}$ their cardinality. For notational convenience, we introduce: $N_\mathcal{C} = N_\mathcal{H} + N_\mathcal{P}$ and $\mathcal{I}_\mathcal{C} = \mathcal{I}_\mathcal{P} \cup \mathcal{I}_\mathcal{H}$ (where $N_\mathcal{C} = N$ and $\mathcal{I}_\mathcal{C} = \{1, 2, \ldots, N\}$).

Given the models (1) for the status and (2)–(3) for the non-discriminant and the discriminant proteins, based on the assumptions that $\boldsymbol{X}_n^+$ and $\boldsymbol{X}_n^-$ are conditionally uncorrelated and that the individual concentrations are also conditionally independent, the distribution of the concentrations and status $(\underline{\boldsymbol{x}}, \boldsymbol{b})$, given the unknown parameters $\boldsymbol{\theta}$ and partition $\boldsymbol{\delta}$, is:

$$\begin{aligned} &f_{\underline{X},\boldsymbol{B}|\boldsymbol{\Theta},\boldsymbol{\Delta}}(\underline{\boldsymbol{x}}, \boldsymbol{b}|\boldsymbol{\theta}, \boldsymbol{\delta}) \\ &\quad = \prod_n f_{X,B|\boldsymbol{\Theta},\boldsymbol{\Delta}}\left(\boldsymbol{x}_n, b_n|\boldsymbol{\theta}, \boldsymbol{\delta}\right) \\ &\quad = \prod_n f_{X|B,\boldsymbol{\Theta},\boldsymbol{\Delta}}\left(\boldsymbol{x}_n|b_n, \boldsymbol{\theta}, \boldsymbol{\delta}\right)\, \mathbb{P}_{B|\boldsymbol{\Theta}}\left(b_n|\boldsymbol{\theta}\right) \\ &\quad = \prod_{n\in\mathcal{I}_\mathcal{P}} \mathcal{N}\left(\boldsymbol{x}_n^+; \boldsymbol{m}_\mathcal{P}, \boldsymbol{\Gamma}_\mathcal{P}\right) \\ &\qquad \times \prod_{n\in\mathcal{I}_\mathcal{H}} \mathcal{N}\left(\boldsymbol{x}_n^+; \boldsymbol{m}_\mathcal{H}, \boldsymbol{\Gamma}_\mathcal{H}\right) \prod_{n\in\mathcal{I}_\mathcal{C}} \mathcal{N}\left(\boldsymbol{x}_n^-; \boldsymbol{m}_\mathcal{C}, \boldsymbol{\Gamma}_\mathcal{C}\right) \\ &\qquad \times \prod_{n\in\mathcal{I}_\mathcal{C}} \mathbb{P}_{B|\boldsymbol{\Theta}}(b_n|\boldsymbol{\theta}) \end{aligned} \tag{4}$$

It can be seen (see Appendix 1) that the exponential arguments of the Gaussian distributions in the first three factors can be reformulated based on the empirical means and covariances for each index set $\mathcal{I}_\mathcal{P}$, $\mathcal{I}_\mathcal{H}$, and $\mathcal{I}_\mathcal{C}$. The result is shown here only for the first factor:

$$\begin{aligned} \prod_{n\in\mathcal{I}_\mathcal{P}} \mathcal{N}\left(\boldsymbol{x}_n^+; \boldsymbol{m}_\mathcal{P}, \boldsymbol{\Gamma}_\mathcal{P}\right) &= (2\pi)^{-P^+ N_\mathcal{P}/2} |\boldsymbol{\Gamma}_\mathcal{P}|^{N/2} \\ \exp\left[-\frac{N_\mathcal{P}}{2}\mathrm{Tr}\left(\boldsymbol{\Gamma}_\mathcal{P}\left[\bar{\mathbf{R}}_\mathcal{P}^+ + \left(\bar{\boldsymbol{x}}_\mathcal{P}^+ - \boldsymbol{m}_\mathcal{P}\right)\left(\bar{\boldsymbol{x}}_\mathcal{P}^+ - \boldsymbol{m}_\mathcal{P}\right)^\mathrm{t}\right]\right)\right] \end{aligned} \tag{5}$$

where $\bar{\boldsymbol{x}}_\mathcal{P}^+$ and $\bar{\mathbf{R}}_\mathcal{P}^+$ are the empirical mean and covariance of the $\boldsymbol{x}_n^+$ for the individuals $n \in \mathcal{I}_\mathcal{P}$. Moreover, regarding the probability for the statuses, we have:

$$\prod_{n\in\mathcal{I}_\mathcal{C}} \mathbb{P}_{B|\boldsymbol{\Theta}}(b_n|\boldsymbol{\theta}) = p^{N_\mathcal{P}}(1-p)^{N_\mathcal{H}}$$

that is only based on the size of each index set $\mathcal{I}_\mathcal{P}$ and $\mathcal{I}_\mathcal{H}$.

Observations Given the previously described concentrations, the proposed developments include two cases for the observation model:

1. In the first one, the concentrations $\boldsymbol{x}_n$ are directly observed.
2. The second one accounts for noise: observations write $\boldsymbol{y}_n = \boldsymbol{x}_n + \boldsymbol{\varepsilon}_n$, where $\boldsymbol{\varepsilon}_n$ is modeled as a zero-mean Gaussian vector with precision $\boldsymbol{\Gamma}_\varepsilon$.

Both of them account for biological variabilities and the latter also includes technological variabilities that arise from both the functioning of the measurement system itself and the post-processing of the spectra. These models are referred to as "noiseless model" and "noisy model". The corresponding variable selection methods are respectively presented in Sections 3.1 and 3.2.

Prior distributions The choice of the prior distribution for the unknown parameters is important. First of all, it must allow us to account for available information

(e.g., nominal values and uncertainties, strong uncertainties,...). Second, it should also enable analytical calculations or numerical computations. To this end, the prior density is chosen as a separable prior distribution, i.e., for the noiseless case

$$\begin{aligned}\pi_{\boldsymbol{\Theta}|\boldsymbol{\Delta}}(\boldsymbol{\theta}|\boldsymbol{\delta}) =& \pi_{\mathcal{P}}\left(\boldsymbol{m}_{\mathcal{P}}, \boldsymbol{\Gamma}_{\mathcal{P}}|\boldsymbol{\delta}\right)\ \pi_{\mathcal{H}}\left(\boldsymbol{m}_{\mathcal{H}}, \boldsymbol{\Gamma}_{\mathcal{H}}|\boldsymbol{\delta}\right)\\ & \pi_{\mathcal{C}}\left(\boldsymbol{m}_{\mathcal{C}}, \boldsymbol{\Gamma}_{\mathcal{C}}|\boldsymbol{\delta}\right)\ \pi_P(p)\end{aligned} \tag{6}$$

for $(\boldsymbol{m}_{\mathcal{P}}, \boldsymbol{\Gamma}_{\mathcal{P}})$, $(\boldsymbol{m}_{\mathcal{H}}, \boldsymbol{\Gamma}_{\mathcal{H}})$, $(\boldsymbol{m}_{\mathcal{C}}, \boldsymbol{\Gamma}_{\mathcal{C}})$, and p. For the noisy case, in addition, we have a factor $\pi(\boldsymbol{\Gamma}_{\varepsilon})$ for $\boldsymbol{\Gamma}_{\varepsilon}$ independent from the other variables. Regarding these five variables individually, the choice is driven by a conjugation principle [30]:

- The probability p is assumed a Beta distributed variable with parameter (a, b).
- The $(\boldsymbol{m}_{\times}, \boldsymbol{\Gamma}_{\times})$ are assumed to be Normal-Wishart $\mathcal{NW}$ distributed with parameters $(\boldsymbol{\mu}_{\times}, \eta_{\times}, \boldsymbol{\Lambda}_{\times}, \nu_{\times})$, for $\times \in \{\mathcal{P}, \mathcal{H}, \mathcal{C}\}$. See Appendix 2.
- The precision $\boldsymbol{\Gamma}_{\varepsilon}$ is under a Wishart distribution with parameters $(\boldsymbol{\Lambda}_{\varepsilon}, \nu_{\varepsilon})$. See Appendix 2.

In the subsequent developments, we proceed with the calculation of the posterior probability for the partitions $\boldsymbol{\delta}$ in the two cases: noiseless concentrations in Section 3.1 and noisy concentrations in Section 3.2. One of the novelty is an explicit analytical result for the noiseless case and a precise approximation for the noisy case.

3.1 Selection using the noiseless data

Optimal decision-maker The question of the paper is the one of the identification of a set of discriminant proteins, and it amounts to making a decision regarding the partition $\boldsymbol{\delta}$. To build an optimal decision-maker, a binary loss is considered that assigns a null loss to the correct decision and a unitary loss to the incorrect decisions. The risk is the mean loss over the models $\boldsymbol{\delta}$, the data $(\underline{\boldsymbol{x}}, \boldsymbol{b})$, and the unknown parameters $\boldsymbol{\theta}$. The optimal decision-maker is defined as the risk minimizer, and it is known to be the one that selects the most a posteriori probable model. It should be noted that alternative loss functions could be chosen, for instance, one that would penalize differently erroneous partitions depending on the number of biomarkers properly identified. In this case, the decision would still be based on the posterior probabilities but with a different rule. However, our choice not only leads to a simple identification procedure but also naturally prevents overfitting.

Thus, the point is to calculate the posterior probability $\mathbb{P}_{\boldsymbol{\Delta}|\underline{\boldsymbol{X}},\boldsymbol{B}}(\boldsymbol{\delta}|\underline{\boldsymbol{x}}, \boldsymbol{b})$ for each candidate model $\boldsymbol{\delta}$. It is carried out using the Bayes rule as:

$$\mathbb{P}_{\boldsymbol{\Delta}|\underline{\boldsymbol{X}},\boldsymbol{B}}(\boldsymbol{\delta}|\underline{\boldsymbol{x}}, \boldsymbol{b}) = \frac{\mathbb{P}(\boldsymbol{\Delta} = \boldsymbol{\delta})\, f_{\underline{\boldsymbol{X}},\boldsymbol{B}|\boldsymbol{\Delta}}(\underline{\boldsymbol{x}}, \boldsymbol{b}|\boldsymbol{\delta})}{\sum_{\boldsymbol{\delta}} \mathbb{P}(\boldsymbol{\Delta} = \boldsymbol{\delta})\, f_{\underline{\boldsymbol{X}},\boldsymbol{B}|\boldsymbol{\Delta}}(\underline{\boldsymbol{x}}, \boldsymbol{b}|\boldsymbol{\delta})} \tag{7}$$

and it crucially depends on the so-called evidence

$$f_{\underline{\boldsymbol{X}},\boldsymbol{B}|\boldsymbol{\Delta}}(\underline{\boldsymbol{x}}, \boldsymbol{b}|\boldsymbol{\delta}) = \int_{\boldsymbol{\theta}} f_{\underline{\boldsymbol{X}},\boldsymbol{B},\boldsymbol{\Theta}|\boldsymbol{\Delta}}(\underline{\boldsymbol{x}}, \boldsymbol{b}, \boldsymbol{\theta}|\boldsymbol{\delta})\ \mathrm{d}\boldsymbol{\theta}$$

which can be rewritten by factorization

$$f_{\underline{\boldsymbol{X}},\boldsymbol{B}|\boldsymbol{\Delta}}(\underline{\boldsymbol{x}}, \boldsymbol{b}|\boldsymbol{\delta}) = \int_{\boldsymbol{\theta}} \pi_{\boldsymbol{\Theta}|\boldsymbol{\Delta}}(\boldsymbol{\theta}|\boldsymbol{\delta})\, f_{\underline{\boldsymbol{X}},\boldsymbol{B}|\boldsymbol{\Theta},\boldsymbol{\Delta}}(\underline{\boldsymbol{x}}, \boldsymbol{b}|\boldsymbol{\theta}, \boldsymbol{\delta})\ \mathrm{d}\boldsymbol{\theta}\,. \tag{8}$$

This calculation is the main difficulty of the paper and more generally in variable and model selection.

In order to carry out this calculation, let us note that the likelihood $f_{\underline{\boldsymbol{X}},\boldsymbol{B}|\boldsymbol{\Theta},\boldsymbol{\Delta}}(\underline{\boldsymbol{x}}, \boldsymbol{b}|\boldsymbol{\theta}, \boldsymbol{\delta})$ factorizes (see Eq. (4)) and that the prior $\pi_{\boldsymbol{\Theta}|\boldsymbol{\Delta}}(\boldsymbol{\theta}|\boldsymbol{\delta})$ also factorizes (see Eq. (6)). So, we have:

$$\begin{aligned}f_{\underline{\boldsymbol{X}},\boldsymbol{B}|\boldsymbol{\Delta}}(\underline{\boldsymbol{x}}, \boldsymbol{b}|\boldsymbol{\delta}) = \int_{\boldsymbol{\theta}} & \mathcal{NW}(\boldsymbol{m}_{\mathcal{P}}, \boldsymbol{\Gamma}_{\mathcal{P}})\, \mathcal{NW}(\boldsymbol{m}_{\mathcal{H}}, \boldsymbol{\Gamma}_{\mathcal{H}})\\ & \mathcal{NW}(\boldsymbol{m}_{\mathcal{C}}, \boldsymbol{\Gamma}_{\mathcal{C}})\ \pi_P(p)\\ & \prod_{n\in\mathcal{I}_{\mathcal{P}}} \mathcal{N}\left(\boldsymbol{x}_n^{+}; \boldsymbol{m}_{\mathcal{P}}, \boldsymbol{\Gamma}_{\mathcal{P}}\right) \prod_{n\in\mathcal{I}_{\mathcal{H}}} \mathcal{N}\left(\boldsymbol{x}_n^{+}; \boldsymbol{m}_{\mathcal{H}}, \boldsymbol{\Gamma}_{\mathcal{H}}\right)\\ & \prod_{n\in\mathcal{I}_{\mathcal{C}}} \mathcal{N}\left(\boldsymbol{x}_n^{-}; \boldsymbol{m}_{\mathcal{C}}, \boldsymbol{\Gamma}_{\mathcal{C}}\right) \prod_{n\in\mathcal{I}_{\mathcal{C}}} \mathbb{P}_{B|\boldsymbol{\Theta}}(b_n|\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}\end{aligned}$$

that can itself be factorized in four integrals: three w.r.t. the couple of variable regarding the concentrations $(\boldsymbol{m}_{\mathcal{P}}, \boldsymbol{\Gamma}_{\mathcal{P}})$, $(\boldsymbol{m}_{\mathcal{H}}, \boldsymbol{\Gamma}_{\mathcal{H}})$, $(\boldsymbol{m}_{\mathcal{C}}, \boldsymbol{\Gamma}_{\mathcal{C}})$ and one w.r.t. the prevalence p:

$$f_{\underline{\boldsymbol{X}},\boldsymbol{B}|\boldsymbol{\Delta}}(\underline{\boldsymbol{x}}, \boldsymbol{b}|\boldsymbol{\delta}) = \mathcal{I}_{\mathcal{P}}^{+}(\underline{\boldsymbol{x}})\, \mathcal{I}_{\mathcal{H}}^{+}(\underline{\boldsymbol{x}})\, \mathcal{I}_{\mathcal{C}}^{-}(\underline{\boldsymbol{x}})\, \mathcal{J}(\boldsymbol{b})$$

where the integrals w.r.t. the $\boldsymbol{m}_{\times}, \boldsymbol{\Gamma}_{\times}$ read

$$\begin{aligned}\mathcal{I}_{\times}^{\star}(\underline{\boldsymbol{x}}) = \int_{\boldsymbol{m}_{\times},\boldsymbol{\Gamma}_{\times}} & \mathcal{NW}(\boldsymbol{m}_{\times}, \boldsymbol{\Gamma}_{\times}) \prod_{n\in\mathcal{I}_{\times}} \mathcal{N}(\boldsymbol{x}_n^{\star}; \boldsymbol{m}_{\times}, \Gamma_{\times})\\ & \mathrm{d}\boldsymbol{m}_{\times}\, \mathrm{d}\boldsymbol{\Gamma}_{\times}\,,\end{aligned} \tag{9}$$

with $\times \in \{\mathcal{P}, \mathcal{H}, \mathcal{C}\}$ and $\star \in \{+, -\}$ and the integral w.r.t. p reads

$$\mathcal{J}(\boldsymbol{b}) = \int_p \pi_P(p) \prod_{n\in\mathcal{I}_{\mathcal{C}}} \mathbb{P}_{\boldsymbol{B}|\boldsymbol{\Theta}}(b_n|\boldsymbol{\theta})\ \mathrm{d}p\,.$$

As far as the three integrals $\mathcal{I}_{\times}^{\star}$ are concerned, the calculations are founded on the reduced form (5) for the likelihood (including empirical means and variances) and on the Normal-Wishart prior (23) in Appendix 2 for the $(\boldsymbol{m}_{\times}, \boldsymbol{\Gamma}_{\times})$. Practically, thanks to the conjugacy property,

the integrand in (9) involves the posterior for $(\boldsymbol{m}_\times, \boldsymbol{\Gamma}_\times)$ which is Normal-Wishart with parameters

$$\begin{aligned}
\nu_\times^{\star\text{pst}} &= \nu_\times + N_\times \\
\eta_\times^{\star\text{pst}} &= \eta_\times + N_\times \\
\boldsymbol{\mu}_\times^{\star\text{pst}} &= \left(N_\times \bar{\boldsymbol{x}}_\times^\star + \eta_\times \boldsymbol{\mu}_\times\right)/(N_\times + \eta_\times) \\
\left(\boldsymbol{\Lambda}_\times^{\star\text{pst}}\right)^{-1} &= (\boldsymbol{\Lambda}_\times)^{-1} + N_\times \bar{\mathbf{R}}_\times^\star + N_\times \eta_\times \left(\boldsymbol{\mu}_\times - \bar{\boldsymbol{x}}_\times^\star\right) \\
&\quad \times \left(\boldsymbol{\mu}_\times - \bar{\boldsymbol{x}}_\times^\star\right)^{\text{t}} / \left(N_\times + \eta_\times\right)
\end{aligned}$$

where "pst" stands for posterior. The finalization of the development relies on the fact that the Normal-Wishart density sums to one: without effective complicate calculus, this yields the result as the ratio of normaliation constants

$$\mathcal{I}_\times^\star = \frac{\mathcal{KNW}_\times^{\star\text{pst}}}{\mathcal{KNW}_\times^{\star\text{pri}}}$$

where $\mathcal{KNW}$ is the normalizing constant of the Normal-Wishart density given by (24) in Appendix 2 and where "pri/pst" stands for "prior/posterior".

As a whole, the analytical calculation of the integral in (8) is possible and yields:

$$f_{\underline{\boldsymbol{X}},\boldsymbol{B}|\boldsymbol{\Delta}}(\underline{\boldsymbol{x}}, \boldsymbol{b}|\boldsymbol{\delta}) \propto \frac{\mathcal{KNW}_\mathcal{P}^{+\text{pst}}}{\mathcal{KNW}_\mathcal{P}^{+\text{pri}}} \frac{\mathcal{KNW}_\mathcal{H}^{+\text{pst}}}{\mathcal{KNW}_\mathcal{H}^{+\text{pri}}} \frac{\mathcal{KNW}_\mathcal{C}^{-\text{pst}}}{\mathcal{KNW}_\mathcal{C}^{-\text{pri}}} \tag{10}$$

rendering the usually complex calculations of the evidences straightforward.

Assuming that all candidate models are equally a priori probable, from Eq. (7), the posterior probability across the 2^P models can be inferred. The selected model is the one which maximizes this probability. It should be noted that if prior information is available such as protein-to-protein interactions (PPI's), it can be taken into account by assigning a higher probability to partitions wherein the related proteins are in the same subset (either discriminant or not). In Eq. (10), the normalizing constants for the posterior distributions depend on the empirical covariance matrices of the population of individuals for the discriminant proteins and the non-discriminant ones, respectively. Their computation is expensive. However, it suffices to compute once the full covariance matrix for all the proteins and then remove the appropriate raws and columns for the 2^P configurations to be tested.

3.2 Selection using noisy data

The model presented above assumes that the concentrations are directly observed. Although this assumption leads to closed-form expressions of the posterior probabilities, it may be too simplifying. In practice, the concentrations are known up to an uncertainty and this section extends the above-detailed developments to account for these uncertainties. However, this comes at the price of intractable calculations, and to overcome this difficulty, we propose a suitable approximation. As introduced above, the measured concentrations are modeled as $\boldsymbol{y}_n = \boldsymbol{x}_n + \boldsymbol{\varepsilon}_n$ where $\boldsymbol{\varepsilon}_n$ is a zero-mean Gaussian vector with precision $\boldsymbol{\Gamma}_\varepsilon$, therefore $\boldsymbol{y}_n|\boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{x}_n; \mathbf{0}, \boldsymbol{\Gamma}_\varepsilon)$. Similarly to the previous section, the vectors of observed concentrations are stacked in a matrix $\underline{\boldsymbol{y}}$ of dimension $P \times N$. To select the most probable model, the evidence $f_{\underline{\boldsymbol{Y}},\boldsymbol{B}|\boldsymbol{\Delta}}(\underline{\boldsymbol{y}}, \boldsymbol{b}|\boldsymbol{\delta})$ must be calculated for each candidate model (it was $f_{\underline{\boldsymbol{X}},\boldsymbol{B}|\boldsymbol{\Delta}}(\underline{\boldsymbol{x}}, \boldsymbol{b}|\boldsymbol{\delta})$ for the noiseless model). The difficulty is that the calculation of evidence requires not only the marginalization of the model parameters but also of the true concentrations. Furthermore, the precision $\boldsymbol{\Gamma}_\varepsilon$ is assumed unknown and must also be marginalized. For notational convenience, we state: $\tilde{\boldsymbol{\theta}} = [\boldsymbol{\theta}, \boldsymbol{\Gamma}_\varepsilon]$ as an extended vector of unknown parameters.

By taking into account the conditional independencies, the evidence can be expressed as:

$$\begin{aligned}
f_{\underline{\boldsymbol{Y}},\boldsymbol{B}|\boldsymbol{\Delta}}(\underline{\boldsymbol{y}}, \boldsymbol{b}|\boldsymbol{\delta}) = \int_{\underline{\boldsymbol{x}}} \int_{\tilde{\boldsymbol{\theta}}} & \prod_{n \in \mathcal{I}_\mathcal{C}} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}_n, \boldsymbol{\Gamma}_\varepsilon) \\
& \prod_{n \in \mathcal{I}_\mathcal{P}} \mathcal{N}\left(\boldsymbol{x}_n^+; \boldsymbol{m}_\mathcal{P}, \boldsymbol{\Gamma}_\mathcal{P}\right) \prod_{n \in \mathcal{I}_\mathcal{H}} \mathcal{N}\left(\boldsymbol{x}_n^+; \boldsymbol{m}_\mathcal{H}, \boldsymbol{\Gamma}_\mathcal{H}\right) \\
& \prod_{n \in \mathcal{I}_\mathcal{C}} \mathcal{N}\left(\boldsymbol{x}_n^-; \boldsymbol{m}_\mathcal{C}, \Gamma_\mathcal{C}\right) \prod_{n \in \mathcal{I}_\mathcal{C}} \mathbb{P}_{B|\boldsymbol{\Theta}}(b_n|\boldsymbol{\theta}) \\
& \pi_{\tilde{\boldsymbol{\Theta}}}\left(\tilde{\boldsymbol{\theta}}|\boldsymbol{\delta}\right) \, \mathrm{d}\tilde{\boldsymbol{\theta}} \, \mathrm{d}\underline{\boldsymbol{x}}
\end{aligned} \tag{11}$$

This multiple integral can be handled in several manners. We propose to first perform integration with respect to $\tilde{\boldsymbol{\theta}}$ and then to integrate the result with respect to $\underline{\boldsymbol{x}}$; hence, Eq. (11) can be rewritten as:

$$f_{\underline{\boldsymbol{Y}},\boldsymbol{B}|\boldsymbol{\Delta}}(\underline{\boldsymbol{y}}, \boldsymbol{b}|\boldsymbol{\delta}) = \int_{\underline{\boldsymbol{x}}} \mathcal{I}_\varepsilon(\underline{\boldsymbol{x}}) \, \mathcal{I}_\mathcal{P}^+(\underline{\boldsymbol{x}}) \, \mathcal{I}_\mathcal{H}^+(\underline{\boldsymbol{x}}) \, \mathcal{I}_\mathcal{C}^-(\underline{\boldsymbol{x}}) \, \mathrm{d}\underline{\boldsymbol{x}}$$

where

$$\mathcal{I}_\varepsilon(\underline{\boldsymbol{x}}) = \int_{\boldsymbol{\Gamma}_\varepsilon} \mathcal{W}(\boldsymbol{\Gamma}_\varepsilon) \prod_{n \in \mathcal{I}_\mathcal{C}} \mathcal{N}(\boldsymbol{y}_n; \boldsymbol{x}_n, \boldsymbol{\Gamma}_\varepsilon) \, \mathrm{d}\boldsymbol{\Gamma}_\varepsilon \tag{12}$$

and in the same manner as previously

$$\begin{aligned}
\mathcal{I}_\times^\star(\underline{\boldsymbol{x}}) = \int_{\boldsymbol{m}_\times, \boldsymbol{\Gamma}_\times} & \mathcal{NW}(\boldsymbol{m}_\times, \boldsymbol{\Gamma}_\times) \prod_{n \in \mathcal{I}_\times} \mathcal{N}\left(\boldsymbol{x}_n^\star; \boldsymbol{m}_\times, \Gamma_\times\right) \\
& \mathrm{d}\boldsymbol{m}_\times \, \mathrm{d}\boldsymbol{\Gamma}_\times
\end{aligned} \tag{13}$$

with $\times \in \{\mathcal{P}, \mathcal{H}, \mathcal{C}\}$ and $\star \in \{+, -\}$. The integrals (12) and (13) can be calculated analytically.

On the one hand, the integrand of (12) can be rewritten, up to a proportionality constant, as the distribution of the precision matrix $\boldsymbol{\Gamma}_\varepsilon$ conditionally upon $\underline{\boldsymbol{y}}$ and $\underline{\boldsymbol{x}}$.

This distribution is Wishart with parameters $\left(\nu_\varepsilon^{\text{pst}}, \boldsymbol{\Lambda}_\varepsilon^{\text{pst}}\right)$ expressed as:

$$\begin{cases} \nu_\varepsilon^{\text{pst}} &= \nu_\varepsilon^{\text{pri}} + N_{\mathcal{C}} \\ \left(\boldsymbol{\Lambda}_\varepsilon^{\text{pst}}\right)^{-1} &= \left(\boldsymbol{\Lambda}_\varepsilon^{\text{pri}}\right)^{-1} + \sum_{n=1}^{N_{\mathcal{C}}} (\boldsymbol{y}_n - \boldsymbol{x}_n)(\boldsymbol{y}_n - \boldsymbol{x}_n)^{\text{t}}. \end{cases}$$

Then, (12) can be re-arranged as:

$$\begin{aligned} \mathcal{I}_\varepsilon(\underline{\boldsymbol{x}}) &= \frac{\mathcal{KW}_\varepsilon^{\text{pst}}}{\mathcal{KW}_\varepsilon^{\text{pri}}} (2\pi)^{-N_{\mathcal{C}}P/2} \\ &= \mathcal{T}_{P,N}\left(\underline{\boldsymbol{x}}; \nu_\varepsilon^{\text{pri}} + 1 - P, \underline{\boldsymbol{y}}, \left(\boldsymbol{\Lambda}_\varepsilon^{\text{pri}}\right)^{-1}, \boldsymbol{I}_N\right) \end{aligned} \tag{14}$$

where $\mathcal{KW}_\varepsilon^\star$ is the normalization constant of the Wishart distribution and $\mathcal{T}_{P,N}(\boldsymbol{T}; q, \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$ denote the matrix t-distribution of parameters q, $\boldsymbol{M}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Omega}$, for a matrix $\boldsymbol{T}$ of dimensions $P \times N$. The expression is recalled in the "Matrix variate t-distribution" section of Appendix 2.

On the other hand, the integrals $\mathcal{I}_\times^\star$ can be computed in the same manner as in the previous section using the conjugation property for the couples $(\boldsymbol{m}_\times, \boldsymbol{\Gamma}_\times)$. Thus, we have:

$$\mathcal{I}_\times^\star(\underline{\boldsymbol{x}}) = (2\pi)^{-N_\times P^\star/2} \ \frac{\mathcal{KNW}_\times^{\star\text{pst}}}{\mathcal{KNW}_\times^{\star\text{pri}}} \tag{15}$$

with $\mathcal{KNW}_\times^{\star\text{pst}}$ and $\mathcal{KNW}_\times^{\star\text{pri}}$ the normalization constants of the prior and posterior Normal-Wishart distributions for $(\boldsymbol{m}_\times, \boldsymbol{\Gamma}_\times)$, respectively.

In (11), the result of the first integration with respect to $\tilde{\boldsymbol{\theta}}$ does not yield an expression that can be integrated analytically w.r.t. $\underline{\boldsymbol{x}}$. To address this issue, we propose to take advantage of the fact that a matrix variate t-distribution $\mathcal{T}_{P,N}(\boldsymbol{T}; q, \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$ tends to a Gaussian distribution when the degrees of freedom parameter q tends to infinity.

In a first step, for a high enough value of $\nu_\varepsilon^{\text{pri}}$, (14) can be approximated as:

$$\mathcal{I}_\varepsilon(\underline{\boldsymbol{x}}) \simeq \prod_{n \in \mathcal{I}_{\mathcal{C}}} \mathcal{N}\left(\boldsymbol{x}_n; \boldsymbol{y}_n, \boldsymbol{\Lambda}_\varepsilon^{\text{pri}} \left(\nu_\varepsilon^{\text{pri}} + 1 - P\right)\right).$$

In a second step, the integrals $\mathcal{I}_\times^\star$ can also be approximated by Gaussian distributions although not directly. For this purpose, we focus on the factor of (15) that depends on the true concentration vectors:

$$\begin{aligned} \mathcal{I}_\times^\star(\underline{\boldsymbol{x}}) = C_\times^\star \Big| &\left(\boldsymbol{\Lambda}_\times^\star\right)^{-1} + N_\times \bar{\mathbf{R}}_\times^\star + N_\times \eta_\times \left(\boldsymbol{\mu}_\times - \bar{\boldsymbol{x}}_\times^\star\right) \\ &\times \left(\boldsymbol{\mu}_\times - \bar{\boldsymbol{x}}_\times^\star\right)^{\text{t}} / \left(N_\times + \eta_\times\right) \Big|^{-\nu_\times^{\text{pst}}/2} \end{aligned} \tag{16}$$

where $C_\times^\star$ is a proportionality constant.

Contrary to (14), the expression (16) does not correspond to a standard probability density function. To make the calculations tractable, we propose to replace the empirical means of the true concentrations $\bar{\boldsymbol{x}}_\times^\star$ by their approximated values computed from the measured concentrations $\bar{\boldsymbol{y}}_\times^\star$. By developing the expression of the empirical covariance matrix in (16), it ensues:

$$\begin{aligned} \mathcal{I}_\times^\star(\underline{\boldsymbol{x}}) &\simeq C_\times^\star \left| \boldsymbol{\Pi}_\times^\star + \sum_{n \in \mathcal{I}_\times} \left(\boldsymbol{x}_n^\star - \bar{\boldsymbol{y}}_\times^\star\right)\left(\boldsymbol{x}_n^\star - \bar{\boldsymbol{y}}_\times^\star\right)^{\text{t}} \right|^{-\nu_\times^{\text{pst}}/2} \\ &= C_\times^\star \left|\boldsymbol{\Pi}_\times^\star\right|^{-\nu_\times^{\text{pst}}/2} \Big| \boldsymbol{I}_P + \left(\boldsymbol{\Pi}_\times^\star\right)^{-1} \left(\underline{\boldsymbol{x}}_\times^\star - \underline{\bar{\boldsymbol{y}}}_\times^\star\right) \\ &\qquad \times \left(\underline{\boldsymbol{x}}_\times^\star - \underline{\bar{\boldsymbol{y}}}_\times^\star\right) \Big|^{-\nu_\times^{\text{pst}}/2}, \\ &\propto \mathcal{T}_{P,N}\left(\underline{\boldsymbol{x}}_\times^\star; \nu_\times^{\star\text{pst}} + 1 - P^\star, \underline{\bar{\boldsymbol{y}}}_\times^\star, \boldsymbol{\Pi}_\times^\star, \boldsymbol{I}_N\right) \end{aligned}$$

where $\boldsymbol{\Pi}_\times^\star = \left(\boldsymbol{\Lambda}_\times^\star\right)^{-1} + N_\times \eta_\times \left(\boldsymbol{\mu}_\times - \bar{\boldsymbol{x}}_\times^\star\right)\left(\boldsymbol{\mu}_\times - \bar{\boldsymbol{x}}_\times^\star\right)^{\text{t}} (N_\times + \eta_\times)$ and $\underline{\bar{\boldsymbol{y}}}_\times^\star$ is a matrix of dimensions $P^\star \times N_\times$ whose columns are all equal to the empirical mean $\bar{\boldsymbol{y}}_\times^\star$. Then, provided $\nu_\times^{\text{pri}}$ is high enough, we can also approximate this matrix variate t-distribution by a Gaussian distribution as for (12):

$$\mathcal{I}_\times^\star(\underline{\boldsymbol{x}}) \propto \prod\nolimits_{n \in \mathcal{I}_\times} \mathcal{N}\left(\boldsymbol{x}_n^\star; \bar{\boldsymbol{y}}_\times^\star, \boldsymbol{\Pi}_\times^\star / \left(\nu_\times^{\text{pri}} - P^\star + 1\right)\right)$$

The performed approximations allow us to express the integrand of the evidence (11) as a product of Gaussian distributions for the true concentration vectors. In this case, the integration can be carried out analytically. By treating separately pathological and healthy individuals, we finally obtain the following expression of the evidence:

$$\begin{aligned} f_{\underline{\boldsymbol{Y}}, \boldsymbol{B}|\boldsymbol{\Delta}}(\underline{\boldsymbol{y}}, \boldsymbol{b}|\boldsymbol{\delta}) \simeq &\left(\frac{\eta_{\mathcal{P}}^{+\text{pri}} \eta_{\mathcal{H}}^{+\text{pri}}}{\eta_{\mathcal{P}}^{\text{pst}} \eta_{\mathcal{H}}^{\text{pst}}}\right)^{P^+} \left(\frac{\eta_{\mathcal{C}}^{+\text{pri}}}{\eta_{\mathcal{C}}^{\text{pst}}}\right)^{P^-} \\ &\times \left|\boldsymbol{\Lambda}_{\mathcal{P}}^{\text{pri}} \boldsymbol{\Pi}_p^+\right|^{-\nu_{\mathcal{P}}^{\text{pri}}} \left|\boldsymbol{\Lambda}_{\mathcal{H}}^{\text{pri}} \boldsymbol{\Pi}_{\mathcal{H}}^+\right|^{-\nu_{\mathcal{H}}^{\text{pri}}} \\ &\times \left|\boldsymbol{\Lambda}_{\mathcal{C}}^{\text{pri}} \boldsymbol{\Pi}_{\mathcal{C}}^+\right|^{-\nu_{\mathcal{C}}^{\text{pri}}} \\ &\times \left|\boldsymbol{\Sigma}_{\mathcal{P}}\right|^{-1/2} \left|\boldsymbol{\Sigma}_{\mathcal{H}}\right|^{-1/2} \\ &\exp\left(-\text{Tr}\left(\boldsymbol{\Sigma}_{\mathcal{P}}^{-1} \bar{\mathbf{R}}_{\mathcal{P}}^{y} + \boldsymbol{\Sigma}_{\mathcal{H}}^{-1} \bar{\mathbf{R}}_{\mathcal{H}}^{y}\right)/2\right) \end{aligned} \tag{17}$$

In this expression, $\bar{\mathbf{R}}_{\mathcal{P}}^{y}$ and $\bar{\mathbf{R}}_{\mathcal{H}}^{y}$ are the empirical covariance matrices of the measured concentration vectors for the pathological and healthy individuals, respectively. As for $\boldsymbol{\Sigma}_{\mathcal{H}}$ and $\boldsymbol{\Sigma}_{\mathcal{P}}$, they are defined as:

$$\begin{aligned} \boldsymbol{\Sigma}_\times = &\frac{\Lambda_\varepsilon^{-1}}{\nu_\varepsilon^{\text{pri}} + 1 - P} \\ &+ \text{blkdiag}\left[\frac{\boldsymbol{\Pi}_\times^+}{\nu_\times^{\text{pri}} + 1 - P^+}, \frac{\boldsymbol{\Pi}_{\mathcal{C}}}{\nu_{\mathcal{C}}^{\text{pri}} + 1 - P^-}\right] \end{aligned}$$

with $\times \in \{\mathcal{P}, \mathcal{H}\}$ and $\text{blkdiag}(\boldsymbol{A}, \boldsymbol{B})$ the block-diagonal matrix with diagonal elements $\boldsymbol{A}$ and $\boldsymbol{B}$.

4 Numerical evaluation

To assess the performance of the proposed method, we have performed extensive numerical experiments using both simulated and real data. They include comparisons with other methods for biomarker identification, namely:

1. The t test [31], which consists in comparing the means of each protein concentrations between the two cohorts, $\mathcal{H}$ and $\mathcal{P}$. If the null hypothesis, standing for the mean equality, is rejected, then the protein is declared as a biomarker. The type I error, denoted as α, corresponds to the incorrect rejection of a true null hypothesis. Its value is used to set the t test decision threshold. In this paper, it is not necessary to adjust the type I errors to account for multivariate effects. The reason is that, for fair comparison purposes, we directly select the setting that leads to the best performance of the test regarding our criterion. This point is commented in Section 4.1 and Fig. 2.
2. The LASSO method [20], based on a linear regression model in which the explanatory variables are the protein concentrations $\underline{\boldsymbol{x}}$, while the response variables are the biological statuses $\boldsymbol{b}$. The LASSO method estimates the coefficients of the model by minimizing the sum of the squared errors, with a L_1-norm penalty. Then, a protein is selected as a biomarker if the value of the coefficient corresponding to its concentration is different from zero. This method introduces a regularization parameter denoted λ.
3. The Bhattacharyya distance [32] is a measure of similarity between two probability distributions and by extension between two populations of individuals [32]. For two multivariate normal distributions with respective mean and covariance matrix $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, it is given by:

$$D_b = \frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T + \frac{1}{2} \log\left(\frac{\det(\boldsymbol{\Sigma})}{\sqrt{\det(\boldsymbol{\Sigma}_1)\det(\boldsymbol{\Sigma}_2)}}\right)$$

   with $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)/2$.
   In the sequel, the Bhattacharyya distance is calculated for each protein by replacing the true mean and covariance matrix by their empirical estimates. The protein is declared as discriminant if the distance is greater than a fixed threshold denoted t. The algorithm is referred to as Bha-distance.
4. The FOHSIC algorithm as introduced in [33]. It performs feature selection based on the Hilbert-Schmidt Independence Criterion (HSIC). The authors propose an unbiased estimator of HSIC and then, assuming the number of significant features is set a priori, use a forward procedure to select them. In our context, the significant features are the biomarkers.

In this section, we refer to the method from Section 3.1 as the Bayesian Model Selection with Analytical Solution for Noiseless Data (BMS-AS-D) method, while to the method from Section 3.2 as the Bayesian Model Selection with Analytical Solution for Noisy data (BMS-AS-N) method.

Crucial to our approach is the choice of the parameters of the Normal-Wishart densities $(\nu_\times, \eta_\times, \boldsymbol{\mu}_\times, \boldsymbol{\Lambda}_\times)$ for $\times \in \{\mathcal{P}, \mathcal{H}, \mathcal{C}\}$. They are referred to as the hyperparameters since the evidence (10) depends on them. In a non-informative case, the values of these hyperparameters are chosen to be $(0, 0, 0, \infty)$, while the proportionality coefficient in (10) has an undetermined form. As an alternative, to tune the parameters, we propose to use a poorly informative prior based on expert knowledge[1] about the corresponding variables (e.g., the values in μg/ml). To this end, we take advantage of the expression of the prior mean and the covariance for $(\boldsymbol{m}_\times, \boldsymbol{\Gamma}_\times)$ as a function of the hyperparameters:

$$E(\boldsymbol{\Gamma}_\times) = \nu_\times \boldsymbol{\Lambda}_\times \quad (18)$$

$$E(\boldsymbol{m}_\times) = \boldsymbol{\mu}_\times \quad (19)$$

$$V(\boldsymbol{m}_\times) = \boldsymbol{\Lambda}_\times^{-1} \,/\, \left[\eta_\times \left(\nu_\times - P^* - 1\right)\right] \quad (20)$$

$$\operatorname{cov}\left(\Gamma_\times^{i,j}, \Gamma_\times^{k,l}\right) = \nu_\times \left(\Lambda_\times^{il}\Lambda_\times^{jk} + \Lambda_\times^{ik}\Lambda_\times^{jl}\right), \quad (21)$$

where the superscripts i,j denote the entry (i,j) of the matrices, $E(\cdot)$ and $V(\cdot)$ refer to the expectation and the covariance matrix of a vector, respectively, while $\operatorname{cov}(\cdot,\cdot)$ stands for the covariance between two random variables. We also recall that $* \in \left\{^+, ^-, ^{\text{“ ”}}\right\}$ depending whether the discriminant/non-discriminant subsets of proteins are considered or the whole set. As a consequence, the prior parameters $(\nu_\times, \eta_\times, \boldsymbol{\mu}_\times, \boldsymbol{\Lambda}_\times)$ can be calculated from (18) to (21) and substituted in (10). Although our choice of prior is not non-informative in the strict sense, it is vague enough so as not to impact biomarker detection. This issue is investigated in the next subsection.

Finally, to calculate (17) in the noisy case, additional hyperparameters for the Wishart probability density function of the noise precision matrix have to be tuned. They are chosen such that $E(\boldsymbol{\Gamma}_\varepsilon) = \nu_\varepsilon^{\text{pri}} \boldsymbol{\Lambda}_\varepsilon^{\text{pri}}$ and that the elements of the covariance matrix satisfy $\operatorname{cov}\left(\Gamma_\varepsilon^{i,j}, \Gamma_\varepsilon^{k,l}\right) = \nu_\varepsilon^{\text{pri}}\left(\Lambda_\varepsilon^{\text{pri},il}\Lambda_\varepsilon^{\text{pri},jk} + \Lambda_\varepsilon^{\text{pri},ik}\Lambda_\varepsilon^{\text{pri},jl}\right)$. Therefore, by accounting for real-life orders of magnitudes of $\boldsymbol{\Gamma}_\varepsilon$, the prior parameters $\left(\nu_\varepsilon^{\text{pri}}, \boldsymbol{\Lambda}_\varepsilon^{\text{pri}}\right)$ can be calculated and substituted in the probability (10).

In the next sections, we present the results of the numerical evaluations of the proposed methods using both simulated and real data.

4.1 Evaluation using simulated data

4.1.1 Description of the simulated data and performance index

We consider the concentrations of a list of P proteins for a group which comprises $N_{\mathcal{H}}$ healthy and $N_{\mathcal{P}}$ pathological individuals, respectively, with $N_{\mathcal{H}} + N_{\mathcal{P}} = N$. The possible partitions for discriminant/non-discriminant proteins thus amount to 2^P and they are referred to as true models. For each true model, $N_{\mathrm{r}} = 10^5$ data realizations are simulated, hence the total number of realizations equals $N_{\mathrm{r}}\, 2^P$.

On the one hand, the noiseless data comprise the biological statuses b_n and the actual protein concentrations $\boldsymbol{x}_n$ of the N individuals and are generated as follows. The biological statuses are sampled from the Bernoulli distribution of parameter p, where p is assumed Beta distributed of parameters $a = 1$ and $b = 1$, which corresponds to a uniform distribution. The concentrations of the subset of discriminant proteins are generated from the Gaussian distributions, $\mathcal{N}\left(\boldsymbol{x}_n^+; \boldsymbol{m}_{\mathcal{H}}, \boldsymbol{\Gamma}_{\mathcal{H}}\right)$ or $\mathcal{N}\left(\boldsymbol{x}_n^+; \boldsymbol{m}_{\mathcal{P}}, \boldsymbol{\Gamma}_{\mathcal{P}}\right)$, depending on the simulated biological status. The subset of non-discriminant proteins are sampled from $\mathcal{N}\left(\boldsymbol{x}_n^-; \boldsymbol{m}_{\mathcal{C}}, \boldsymbol{\Gamma}_{\mathcal{C}}\right)$. The parameters $(\boldsymbol{m}_\times, \boldsymbol{\Gamma}_\times)$, where $\times \in \{\mathcal{P}, \mathcal{H}, \mathcal{C}\}$, are distributed according to the Normal-Wishart distribution $\mathcal{NW}(\nu_\times, \eta_\times, \boldsymbol{\mu}_\times, \boldsymbol{\Lambda}_\times)$. The orders of magnitudes for $(\boldsymbol{m}_\times, \boldsymbol{\Gamma}_\times)$ are specified as: $E(\boldsymbol{m}_\times) = 10^3 \mathbf{1}_{P^*}$, $V(\boldsymbol{m}_\times) = 10^4\, \mathbf{I}_{P^*}$, $E(\boldsymbol{\Gamma}_\times) = 10^3\, \mathbf{I}_{P^*}$, $V(\boldsymbol{\Gamma}_\times) = 10^4\, \mathbf{I}_{P^*}$, where $\mathbf{1}_{P^*}$ denotes a vector of size P^* whose elements are all equal to 1 and $\mathbf{I}_{P^*}$ is the identity matrix of size P^*. The same order of magnitude is considered for healthy, pathological, and common parameters, that is to say $\times \in \{\mathcal{H}, \mathcal{P}, \mathcal{C}\}$. These a priori information are used to tune the hyperparameters as given by (18)–(21).

On the other hand, the noisy data include the biological statuses b_n and the observed protein concentrations $\boldsymbol{y}_n$ for $n = 1, 2 \ldots N$. The protein concentrations $\boldsymbol{x}_n$ are generated as in the case of the noiseless observations by using the same hyperparameter setting. As for the noise $\boldsymbol{\varepsilon}_n$, it is sampled as a zero-mean multivariate Gaussian random vector with precision matrix $\boldsymbol{\Gamma}_\varepsilon$. The latter is generated from a Wishart density with parameters $\left(\nu_\varepsilon^{\mathrm{pri}}, \boldsymbol{\Lambda}_\varepsilon^{\mathrm{pri}}\right)$. In order to determine these hyperparameters, the a priori information is specified as: $E(\boldsymbol{\Gamma}_\varepsilon) = 10^{-2}\, \mathbf{I}_P$, $V(\boldsymbol{\Gamma}_\varepsilon) = 10^{-5}\, \mathbf{I}_P$. Note here that $\boldsymbol{\Gamma}_\varepsilon$ measures the precision (inverse variance), thus the lower $\boldsymbol{\Gamma}_\varepsilon$ is, the stronger the noise is.

For each data set, the posterior probability is computed for all possible partitions according to (10) or (17) for the BMS-AS-D and BMS-AS-N methods, respectively. Then, the most probable partition is selected.

The performance is measured in terms of the error rate τ, defined as:

$$\tau(\%) = \left(\sum_{i=1}^{2^P} E_i / \left(N_{\mathrm{r}} \times 2^P \right) \right) \times 100$$

where E_i is the number of realizations of the ith partition for which the selected model is different from the true one. It should be noted that the usual type I and type II errors apply when the biomarker identification is made protein after protein but they are not relevant here, since the proteins are addressed as a whole. Indeed, as underlined in Section 3, the proposed Bayesian formulation relies on a different paradigm than the existing methods: any wrong or correct decision regards the list of proteins as a whole (and not protein-wise). Furthermore, the proposed approach minimizes a global risk and cannot be directly related to a given false discovery rate. It ensues that τ, which encompasses all types of errors, appears as a suitable criterion to measure the performance.

4.1.2 Results for the noiseless model

First of all, we investigate the sensitivity of the error rate $\tau(\%)$ to the hyperparameter tuning. In Fig. 1, we plot the latter as a function of the considered variance $V(\boldsymbol{m}_\times)$ in Eqs. (18)–(21). We can observe that provided the variance is chosen superior to a given threshold, it has no impact on the biomarker detection performance. This is the case for our settings.

Before going further in analyzing the performance, it should be noted that the t test, the LASSO, and the Battacharyya distance all require the setting of a parameter: the type I error α, the regularization parameter λ, and the threshold t, respectively. So as not to favor our approach, we have run all the algorithms for different values of these parameters and we have selected the best one (in order to get the lowest error rate). Such a procedure cannot be applied on real data, but it allows us to compare the proposed method to the best version of the alternative approaches. The results are given in Fig. 2 for the t test and the LASSO.

The performance of the BMS-AS-D method has first been evaluated as a function of the number of proteins and results are shown in Table 1. The superior performance

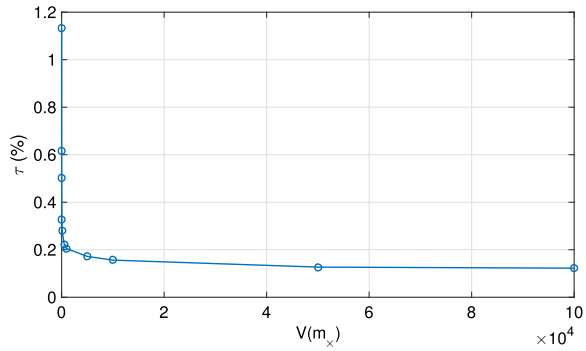

Fig. 1 The mean is fixed, and different values of the variance are considered. $P = 3$, $N = 10^3$, $N_{\mathrm{r}} = 10^5$

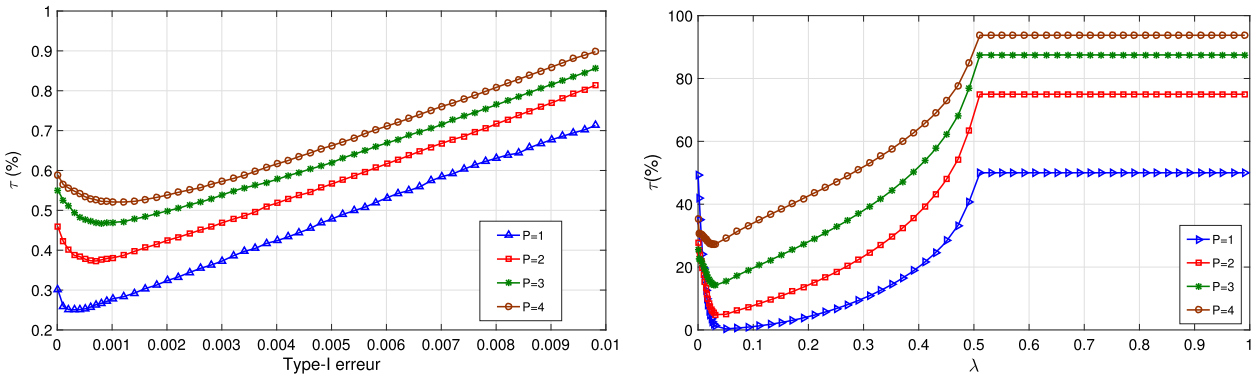

Fig. 2 Error rate τ(%) for the *t* test with different values of the type I error (*left*) and for the LASSO with different values of λ (*right*)

of the BMS-AS-D method with respect to the *t* test is expected since the BMS-AS-D makes a decision jointly across all the proteins while accounting for all possible correlations between them. This issue is not addressed within the *t* test which is univariate, i.e., each protein is tested separately. The same observation can be made for the Bhattacharyya distance. Indeed, the error rate of the *t* test and the Bhattacharyya distance are very close. As for the superiority of the BMS-AS-D method over the LASSO, it is due to the fact that the latter makes the assumption of an arbitrary linear regression relationship between the biological statuses and the protein concentrations. Moreover, the difference in performance increases with the number of proteins, since the possibility of correlation increases. Indeed, in the presence of correlated variables, the number of significant variables is known to be over-estimated with the LASSO algorithm. This fact confirms the relevance of the BMS-AS-D method which can accommodate correlations between the variables.

Regarding the number of individuals, Table 2 shows that, as expected, the performance of all the methods improves as the number of considered individuals increases. In particular, the better performance of the BMS-AS-D method is explained by the reduced variance of the protein concentration posterior distribution for a large number of individuals. Furthermore, the BMS-AS-D method outperforms the *t* test, the LASSO and the Bhattacharyya distance, regardless the number of individuals. Finally, even if the number of configurations to be tested increases exponentially with the number of proteins, the computational cost is kept reasonable owing to the analytical expression of the posterior probabilities for the different partitions.

Last but not the least, our method can also be run for a fixed number of biomarkers as it is the case of many current feature identification algorithms such as the FOHSIC. Only the partitions with the proper number of biomarkers are studied, which amounts to assigning a null prior probability to the others. When the number of biomarkers is chosen as $M \leq P$, the number of posterior probabilities to compute is limited to C_P^M instead of 2^P. Under this assumption, the performance of the BMS-AS-D is compared to that of the FOHSIC in the Tables 3, 4, 5 and 6. To run the algorithm with large P ($\geq$ 8), the number of realizations is reduced to 10^3.

As shown in Tables 3, 5, and 6, the BMS-AS-D algorithm outperforms the FOHSIC one. This is explained by the fact that the BMS-AS-D algorithm makes a multivariate decision on the whole set of proteins, while the FOHSIC uses a forward procedure which can lead to error accumulation. Indeed, if any detected biomarker in the sequence is false, then the final selected model is bound to be erroneous. Furthermore, the BMS-AS-D is also faster than the FOHSIC, as illustrated in Table 4. More precisely, Table 5 shows the error rate τ for the FOHSIC and the Bayesian algorithm for $P = 8$ and different number of biomarkers. Conversely, the results proposed in Table 6 are obtained with the number of biomarkers fixed to $M = 4$ while

Table 1 Noiseless data: τ (%) for different value of P, $N = 1000$

P	1	2	3	4
Best *t* test	0.2555	0.4025	0.471	0.5209
Best LASSO	0.3185	4.865	14.207	27.2655
Best Bha-distance	0.2245	0.3743	0.4496	0.5008
BMS-AS-D	0.0935	0.1434	0.1563	0.1616

Table 2 Noiseless data: τ (%) for different values of N, $P = 3$

N	100	500	1000
t test	4.041	0.8938	0.471
LASSO	20.5541	15.9870	14.207
Best Bha-distance	3.6970	0.8194	0.4496
BMS-AS-D	1.71	0.32	0.1563

Table 3 τ (%) $N = 1000$ and $P = 3$

Number of biomarkers	2
BMS-AS-D	0.0723
FOHSIC	0.2907

Table 5 τ (%) for $P = 8$, $N_r = 10^3$, and $N = 500$

M:	4	6
BMS-AS-D	0.3014	0.2107
FOHSIC	0.8271	0.9107

the number of proteins P is varied. As expected, the performance of the FOHSIC algorithm is degraded when increasing the number of proteins while the opposite is observed for the BMS-AS-D. Thus, even for large P, the Bayesian algorithm outperforms the FOHSIC.

4.1.3 Results for the noisy model

The performance of the BMS-AS-N and the BMS-AS-D algorithms is first studied as a function of the number of proteins P and the number of individuals N, for a fixed noise level. Then, the BMS-AS-N and the BMS-AS-D methods are compared for different noise conditions.

Table 7 reports the error rate τ (%) for the BMS-AS-N and the BMS-AS-D methods. The value of the error rate for the BMS-AS-D method is increased as compared to the results given in Table 1. This is due to the fact that the BMS-AS-D relies on a noiseless model, i.e., it processes the noisy data as if they were noiseless protein concentrations. That is why this result is expected, especially given the specified severe noise conditions $\left(E(\mathbf{\Gamma}_\varepsilon) = 10^{-2}\,\mathbf{I}_P\right)$. As a consequence, the performance of the BMS-AS-N method is significantly better than the BMS-AS-D one. Also, the difference in performance increases with the number of proteins, since for large sets of proteins, the number of candidate models increases and the estimation becomes more difficult.

The performance of the BMS-AS-N method is also evaluated for several numbers of individuals: $N = 100, 500$, and 1000. Table 8 shows the results where it can be observed that for $N = 100$, the error rate is higher; however, it does not exceed 13%. For the BMS-AS-D method, the results are even poorer because of the impact of the noise on smaller sample sizes.

To assess the importance of taking into account the noise in the model, the respective performances of the BMS-AS-D and the BMS-AS-N methods are also compared for different noise levels. We recall that the former is designed from the noiseless data model, while the latter specifically addresses the noisy data model. The noise power is measured by the mean value of the noise variance $E(\mathbf{\Gamma}_\varepsilon)$, which is varied in the simulations. Table 9 shows the error rate for both methods as a function of this parameter: The BMS-AS-N method always outperforms the BMS-AS-D one. In the absence of noise, the BMS-AS-N method becomes equivalent to the optimal noiseless method. These results confirm the relevance of the method, especially for high noise levels.

As a conclusion, the results confirm the good performance of the proposed BMS-AS-N method which is also not too computationally intensive by means of the analytical approximation of the posterior probabilities.

4.2 Evaluation using the real data

The primary goal of this paper was to present a novel methodology for biomarker identification that relaxes classical simplifying assumptions on the data model and then to evaluate it on simulated data. Nevertheless, we had at our disposal a batch of real data[2] and we used it to carry out a preliminary study of the BMS-AS-N method. The data are composed of 206 samples: 105 with the status $\mathcal{H}$ (including 76 patients from blood donors and 29 with negative colonoscopy), 101 with malignant tumor[3], i.e. with status $\mathcal{P}$. The latter are structured as follows: 24 patients in the 'stage one' of the cancer, 26 patients in the 'stage two', 23 patients in the 'stage three', 25 patients in the 'stage four', and three missing values. The protein concentrations are obtained using the Bayesian inversion method developed in [34] from measurements of SRM spectra according to the methodology described in [35]. For each sample, the concentrations of 21 proteins are measured (14-3-3 protein sigma; 78-kDa glucose-regulated protein; protein S100-A11; calmodulin; calreticulin; peptidyl-prolyl cis-trans isomerase A; defensin-5; defensin-6; heat shock cognate 71 kDa protein; fatty acid-binding protein, intestinal; fatty acid-binding protein, liver (LFABP); stress-70 protein, mitochondrial; protein disulfide-isomerase (PDI); protein disulfide-isomerase A6 (PDIA6); phosphoglycerate kinase 1; retinol-binding protein 4; peroxiredoxin 5, mitochondrial; protein S100-A14; triosephosphate isomerase; villin-1 (Villin); Vimentin). Only one of the proteins in the sample, named LFABP, was previously identified by SRM as a biomarker [36]. To

Table 4 Execution time for one simulation $N = 1000$ and $P = 3$

Number of biomarkers	2
BMS-AS-D	0.202 s
FOHSIC	0.732 s

Table 6 τ (%) $M = 4$, $N_r = 10^3$, and $N = 500$

P	8	12
BMS-AS-D	0.3014	0.1818
FOHSIC	0.8271	0.8103

Table 7 Noisy data: τ (%) for different values of P, $N = 1000$

P	1	2	3	4
BMS-AS-D	16.82	32.784	48.740	63.686
BMS-AS-N	1.38	2.682	4.158	5.698

Table 9 τ (%) for $N = 1000$, $P = 3$

$E(\mathbf{\Gamma}_\varepsilon)(\times \mathbf{I}_P)$	10^{-2}	10^{-1}	1	10	10^2
BMS-AS-D	48.740	12.304	3.064	0.813	0.291
BMS-AS-N	4.158	1.395	0.567	0.303	0.220

calculate the hyperparameters (18)–(21), empirical orders of magnitudes for $(\boldsymbol{m}_\times, \mathbf{\Gamma}_\times)$ (e.g., µg/ml) are used as specified: $E(\boldsymbol{m}_\times) = 10^2 \mathbf{1}_{P^*}$, $V(\boldsymbol{m}_\times) = 10^3\, \mathbf{I}_{P^*}$, $E(\mathbf{\Gamma}_\times) = 10^3\, \mathbf{I}_{P^*}$, $V(\mathbf{\Gamma}_\times) = 10\, \mathbf{I}_{P^*}$.

For this data set, the posterior probability is computed for each of the 2^{21} possible partitions according to (17) for the BMS-AS-N methods. Table 10 presents the four most probable partitions, with their probabilities. By far, the most probable partition (probability 0.9986) is: LFAPB is discriminant and the remaining 20 proteins are non-discriminant. The second most probable partition is: the whole set of protein is non-discriminant (probability 0.001361). The third and the fourth ones select two discriminant proteins and the 19 other are non-discriminant: (LFABP , Villin) and (LFABP , PDIA6), with probability smaller than 10^{-6}. As a conclusion, the proteins are all declared as non-discriminant, except the LFAPB. This study confirms that our method correctly identifies the valid biomarker.

Despite the large number of models to compare (about two millions candidate models), the computation time is just 1 h. This short computation time is made possible by the analytical calculation of the posterior probability, avoiding the use of extensive numerical integration methods such as for instance MCMC algorithms [24].

5 Synthesis and perspectives

Biomarker discovery is a challenging task of the utmost interest for the diagnosis and prognosis of diseases. This paper presents a statistical approach based on variable selection. It is developed in a Bayesian framework that relies on an optimal strategy, i.e., the minimization of an error risk. Given P candidate proteins, the proposed procedure compares the probability of the 2^P partitions (subset of discriminant versus subset of non-discriminant proteins). The most a posteriori probable partition is finally retained and thus defines the selected variables. The main difficulty is the required integration with respect to all the unknown model parameters. An important contribution is to provide a closed-form expression of the probabilities for noiseless observations and a sensible approximation for noisy observations. The proposed method proved to be well-suited for variable selection in a complex context. Its effectiveness is assessed by a theoretical characterization and numerical studies (on simulated and real data) which are in accordance with the theoretical optimality. Furthermore, the proposed method compares favorably with the t test, the LASSO, the Battacharrya distance, and the FOHSIC.

From a methodological standpoint, several perspectives can be considered. Regarding the concentrations, non-Gaussian distributions, e.g., Gamma or Wishart models, could be relevant. Regarding the status, a possible development could account for possible errors in the given status. In this case, an additional level should be appended to the hierarchical Bayesian model. It would include a prior probability for an erroneous status.

As for the applicative perspectives, we plan to further take advantage of the performance of the method in other clinical data sets or in other biomedical fields (e.g., genomics, transcriptomics…). In addition, we also intend to make use of the method in other domains, for instance, in astrophysics (identification of pertinent features in order to classify galaxies), or for complex structures and industrial processes (identification of indicators for detection and diagnosis of damages or faults, analysis of fatigue and aging prevention,…).

Table 8 Noisy data: τ (%) for different valued of N, $P = 3$

N	100	500	1000
BMS-AS-D	86.850	71.615	48.740
BMS-AS-N	12.477	5.781	4.158

Endnotes

[1] The knowledge about orders of magnitudes of the concentration values is acquired from the real data set provided by bioMérieux (Technology Research Department), France.

[2] SRM measurements provided by bioMérieux (Technology Research Department), France

[3] colorectal cancer

Appendix 1

Reduction of the concentration distribution

This section explains how the exponential arguments of the Gaussian distributions in (4) can be reformulated

Table 10 The four most probable partitions, for real data $P = 21$

Declared biomarker	LFABP	No biomarker	LFABP and Villin	LFAPB and PDIA6
Probability	9.986×10^{-1}	1.361×10^{-3}	9.762×10^{-7}	8.297×10^{-7}

based on the empirical means and covariances of the concentrations, yielding relation (5). We have:

$$\begin{aligned}\prod_{n\in\mathcal{I}_\mathcal{P}} \mathcal{N}\left(\boldsymbol{x}_n^+;\boldsymbol{m}_\mathcal{P},\boldsymbol{\Gamma}_\mathcal{P}\right) &= \prod_{n\in\mathcal{I}_\mathcal{P}} (2\pi)^{-P^+/2}|\boldsymbol{\Gamma}_\mathcal{P}|^{1/2} \\ &\exp\left[-\left(\boldsymbol{x}_n-\boldsymbol{m}_\mathcal{P}\right)^{\mathrm{t}}\boldsymbol{\Gamma}_\mathcal{P}\left(\boldsymbol{x}_n-\boldsymbol{m}_\mathcal{P}\right)/2\right] \\ &= (2\pi)^{-P^+N_\mathcal{P}/2}|\boldsymbol{\Gamma}_\mathcal{P}|^{N_\mathcal{P}/2} \\ &\exp\left[-\sum_{n\in\mathcal{I}_\mathcal{P}}\left(\boldsymbol{x}_n-\boldsymbol{m}_\mathcal{P}\right)^{\mathrm{t}}\boldsymbol{\Gamma}_\mathcal{P}\left(\boldsymbol{x}_n-\boldsymbol{m}_\mathcal{P}\right)/2\right]\end{aligned}$$

The idea is to re-arrange the sum in the exponential as a function of the empirical mean $\bar{\boldsymbol{x}}_\mathcal{P}^+$ and the empirical variance $\bar{\mathbf{R}}_\mathcal{P}^+$. To this end, for the sake of calculation simplicity, we can write: $\boldsymbol{x}_n-\boldsymbol{m}_\mathcal{P}=\left(\boldsymbol{x}_n-\bar{\boldsymbol{x}}_\mathcal{P}^+\right)-\left(\bar{\boldsymbol{x}}_\mathcal{P}^+-\boldsymbol{m}_\mathcal{P}\right)$ and develop the sum of product. Then, using the fact that $\boldsymbol{u}^{\mathrm{t}}\boldsymbol{M}\boldsymbol{u}=\mathrm{Tr}\left(\boldsymbol{u}^{\mathrm{t}}\boldsymbol{M}\boldsymbol{u}\right)=\mathrm{Tr}\left(\boldsymbol{M}\boldsymbol{u}\boldsymbol{u}^{\mathrm{t}}\right)$, the product is written as function of empirical mean and variance

$$\begin{aligned}\prod_{n\in\mathcal{I}_\mathcal{P}} \mathcal{N}\left(\boldsymbol{x}_n^+;\boldsymbol{m}_\mathcal{P},\boldsymbol{\Gamma}_\mathcal{P}\right) &= (2\pi)^{-P^+N_\mathcal{P}/2}|\boldsymbol{\Gamma}_\mathcal{P}|^{N/2} \\ &\exp\left[-\frac{N_\mathcal{P}}{2}\mathrm{Tr}\left(\boldsymbol{\Gamma}_\mathcal{P}\left[\bar{\mathbf{R}}_\mathcal{P}^+ + \left(\bar{\boldsymbol{x}}_\mathcal{P}^+-\boldsymbol{m}_\mathcal{P}\right)\left(\bar{\boldsymbol{x}}_\mathcal{P}^+-\boldsymbol{m}_\mathcal{P}\right)^{\mathrm{t}}\right]\right)\right]\end{aligned} \tag{22}$$

which allows easier handling of the Normal-Wishart prior.

Appendix 2

Wishart, Normal-Wishart, and matrix variate t-distribution

Wishart

The Wishart density probability function for a $P\times P$ matrix $\boldsymbol{\Gamma}$ is driven by two parameters: a degree of freedom ν (real and larger than $P-1$) and a matrix $\boldsymbol{\Lambda}$ (positive and definite) referred to as the scale matrix. It reads

$$\begin{aligned}\mathcal{W}(\boldsymbol{\Gamma}\,;\,\boldsymbol{\Lambda},\nu) = \mathcal{KW}^{-1}\,&\det\left[\boldsymbol{\Gamma}\right]^{(\nu-P-1)/2} \\ &\exp\left[-\mathrm{Tr}\left[\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\right]/2\right]\end{aligned}$$

The normalizing constant $\mathcal{KW}$ depends on ν and $\boldsymbol{\Lambda}$:

$$\mathcal{KW}=\mathcal{KW}(\nu,\boldsymbol{\Lambda})=2^{\nu P/2}\det\left[\boldsymbol{\Lambda}\right]^{\nu/2}\;\gamma_P(\nu/2)$$

where γ_P is the P-dimensional Gamma function.

Normal-Wishart

For a couple $(\boldsymbol{m},\boldsymbol{\Gamma})$, where $\boldsymbol{m}$ is a P-dimensional vector and $\boldsymbol{\Gamma}$ a $P\times P$ matrix, the Normal-Wishart density is controled by four parameters $\nu,\eta,\boldsymbol{\mu},\boldsymbol{\Lambda}$. It reads:

$$\begin{aligned}\mathcal{NW}(\boldsymbol{m},\boldsymbol{\Gamma}\,;\,\nu,\eta,\boldsymbol{\mu},\boldsymbol{\Lambda}) = \mathcal{KNW}^{-1}\,&\det\left[\boldsymbol{\Gamma}\right]^{(\nu-P)/2} \\ &\exp\left[-\left[\mathrm{Tr}\left[\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\right]\right.\right. \\ &\left.\left.+\eta(\boldsymbol{m}-\boldsymbol{\mu})^{\mathrm{t}}\boldsymbol{\Gamma}(\boldsymbol{m}-\boldsymbol{\mu})\right]/2\right]\end{aligned} \tag{23}$$

and the normalizing constant is:

$$\begin{aligned}\mathcal{KNW}=\mathcal{KNW}(\nu,\eta,\boldsymbol{\mu},\boldsymbol{\Lambda}) = (2\pi)^{P/2}\,2^{\nu P/2}\,\eta^{-P/2} \\ \det\left[\boldsymbol{\Lambda}\right]^{\nu/2}\;\gamma_P(\nu/2)\,,\end{aligned} \tag{24}$$

and it does not depend on $\boldsymbol{\mu}$ (that is a position parameter).

Matrix variate t-distribution

The random matrix $\boldsymbol{T}$ of dimension $(P\times N)$ is said to have a matrix variate t-distribution with parameters $\boldsymbol{M},\boldsymbol{\Sigma},\boldsymbol{\Omega}$, and q if its probability density function is given by

$$\begin{aligned}\mathcal{T}_{P,N}(q,\boldsymbol{M},q\boldsymbol{\Sigma},\boldsymbol{\Omega}) = \;&\frac{\gamma_P\left(\left[q+N+P-1\right]/2\right)}{\gamma_P\left(\left[P+p-1\right]/2\right)}\,\pi^{-NP/2} \\ &|\boldsymbol{\Sigma}|^{-N/2}\,|\boldsymbol{\Omega}|^{-P/2}\,|\boldsymbol{I}_P+\boldsymbol{\Sigma}^{-1}(\boldsymbol{T}-\boldsymbol{M}) \\ &\boldsymbol{\Omega}^{-1}(\boldsymbol{T}-\boldsymbol{M})^{\mathrm{t}}|^{-\left[q+N+P-1\right]/2}\end{aligned}$$

where $\boldsymbol{T}$ and $\boldsymbol{M}$ are $P\times N$ matrices, $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ are positive-definite matrices with respective sizes $N\times N$ and $P\times P$ and $q>0$.

When q tends to infinity, the distribution of $\boldsymbol{T}$ tends to a Gaussian distribution with mean $\boldsymbol{M}$ and covariance $\boldsymbol{\Sigma}\otimes\boldsymbol{\Omega}$ that is to say $\mathcal{N}(\boldsymbol{T}\,;\,\boldsymbol{M},\boldsymbol{\Sigma}\otimes\boldsymbol{\Omega})$, where $\otimes$ is the Kronecker product.

Funding

This work was supported by the French National Research Agency through the Bayesian Hierarchical Inversion in Proteomics Project under Contract ANR-2010-BLAN-0313.

Authors' contributions

JFG proposed the initial idea of the method and provided the first theoritical developments. He also importantly contributed to the writing of the paper. AG proposed the extension to noisy data and importantly contributed to the writing of the paper. ND provided the largest part of the Matlab developments and numerical assessment. She also contributed to the writing of the paper. CT provided input in the proteomics fields. She contributed to the comparison with existing methods. She also contributed to the manuscript and proposed valuable comments. MH contributed to the Matlab development and the numerical assessment. She also contributed to the writing of the paper. JPC provided input in the SRM field and with the result interpretation. He acquired and provided real SRM spectra to test the algorithm. He has revised the manuscript. LG has developed the BHI processing algorithms for the MRM mode. He has computed the protein concentration profiles for the MRM clinical data set used for the evaluation on real data. He contributed to the result interpretation. PD provided expertise regarding proteomics. BL contributed to the early developments of the BHI-PRO project. PG was the BHI-PRO project manager. He has coordinated the conception of the processing algorithms and the interpretation of the results. He has revised the manuscript. PR coordinated biostatistical developments and the interpretation of the results. He has revised the manuscript. All authors read and approved the final manuscript.

Competing interests

JPC and BL are employed by bioMérieux. The other authors declare that they have no competing interests.

Publisher's Note

Author details
[1]IMS (Univ. Bordeaux, CNRS, BINP), 33400 Talence, France. [2]National Engineering School of Gabes (ENIG), University of Gabes, Gabes, Tunisia. [3]CLIPP, Pôle de Recherche Université de Bourgogne, 21000 Dijon, France. [4]NATO STO Centre for Maritime Research and Experimentation, 19126 La Spezia, Italy. [5]Technology Research Department, Innovation Unit, bioMérieux SA, Marcy l'Étoile, France. [6]Univ. Grenoble Alpes, F-38000 Grenoble, France. [7]CEA, LETI, MINATEC Campus, F-38054 Grenoble, France. [8]Service de Biostatistique - Bioinformatique, Hospices Civils de Lyon, Lyon, France. [9]CNRS UMR 5558, LBBE, Équipe Biostatistique Santé, Villeurbanne, France. [10]Université de Lyon, Université Claude Bernard Lyon 1, Lyon, France. [11]Pôle Rhône-Alpes de Bioinformatique, Université Claude Bernard - Lyon 1, 69622 Villeurbanne, France.

Received: 4 August 2016 Accepted: 21 June 2017
Published online: 14 July 2017

References

1. S Srivastava, *Informatics in Proteomics,* ser. Statistics: a series of textbooks and monographs. (CRC Press, Boca Raton, 2005)
2. KA Do, P Muller, M Vannucci, *Bayesian inference for gene expression and proteomics*. (Cambridge University Press, Cambridge, England, 2006)
3. T Fortin, A Salvador, JP Charrier, C Lenz, X Lacoux, A Morla, G Choquet-Kastylevsky, J Lemoine, Clinical quantitation of prostate-specific antigen biomarker in the low nanogram/milliliter range by conventional bore liquid chromatography-tandem mass spectrometry (multiple reaction monitoring) coupling and correlation with ELISA tests multiple hypothesis testing in microarray experiments. Mol. Cell Proteomics. **8**(5), 1006–1015 (2009)
4. C Huillet, A Adrait, D Lebert, G Picard, M Trauchessec, M Louwagie, A Dupuis, L Hittinger, B Ghaleb, P Le Corvoisier, M Jaquinod, J Garin, C Bruley, V Brun, Accurate quantification of cardiovascular biomarkers in serum using protein standard absolute quantification (PSAQ) and selected reaction monitoring. Mol. Cell Proteomics. **11**(2) (2012)
5. K Harris, M Girolami, H Mischak, Definition of valid proteomic biomarkers: a Bayesian solution. Lett. Notes Comput. Sci. **5780**, 137–149 (2009)
6. M Frantzi, A Bhat, A Latosinska, Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. Clin. Transl. Med. **3**(7) (2014)
7. P Roy, C Truntzer, D Maucort-Boulch, T Jouve, N Molinari, Protein mass spectra data analysis for clinical biomarker discovery: a global review. Brief. Bioinform. **12**(2), 176–186 (2011)
8. D Sidransky, S Srivastava, Changes in collagen metabolism in prostate cancer: a host response that may alter progression. Nat. Rev. Cancer. **18**(3), 789–795 (2003)
9. H Hoijtink, I Klugkist, Comparison of hypothesis testing and Bayesian model selection. Qual. Quant. **41**, 73–91 (2007)
10. S Dudoit, J Popper Shaffer, J Boldrick, Multiple hypothesis testing in microarray experiments. Statist. Sci. **18**, 71–103 (2003)
11. Y Benjamin, Y Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B. **57**(1), 289–300 (1995)
12. GK Smyth, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. **3**(3) (2004)
13. N Draper, H Smith, *Applied regression analysis*, 3rd ed. (Wiley Series in Probability and Statistics, Chichester, New York, Singapore, Toronto, 1998)
14. M Bhattacharlee, C Botting, M Sillanpaa, Bayesian biomarker identification based on marker-expression proteomics data. ELSEVIER Genomics. **92**, 37–55 (2008)
15. J Fan, R Li, Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
16. M Chen, Dc Dey, Variable selection for multivariate logistic regression model. J. Stati. Plan. and Infer. **111**, 37–55 (2003)
17. Z Yuan, D Ghosh, Combining multiple biomarker models in logistic regression. Biometrics. **64** (2008)
18. H Akaike, Information theory and an extension of the maximum likelihood principle. Proc. Second Int. Symp. Inform, 261–281 (1973)
19. AE Hoerl, RW Kennard, Ridge regression: applications to nonorthogonal problems. Technometrics. **12**(1), 69–82 (1970)
20. R Tibshirani, Regression shrinkage and selection via the LASSO. J. Royal Stat. Soc,: Series B (Methodology). **1**, 267–288 (1996)
21. H Zou, T Hastie, Regularization and variable selection via the elastic net. J. R. Statist. Soc. B. **67**, 301–320 (2005)
22. J Ogutu, T Schultz-Streek, HP Piepho, Genomic selection using regularized linear regression models: ridge regression, LASSO, elastic net and their extensions. ser. Proc. of the 15th European workshop on QTL mapping and marker assisted selection (QTLMAS), Rennes, France, 2011, 37–55
23. MI Ghahramani, A note on the evidence and Bayesian Occam's razor.Gatsby Unit, University College London, Technical Report GCNU-TR 2005-003 (2005)
24. CP Robert, G Casella, *Monte-Carlo statistical methods*, ser. Springer Texts in Statistics. (Springer, New York, 2004)
25. B Carlin, T Louis, *Bayesian methods for data analysis*. (CRC Press, Chapman & Hall, Boca Raton, London, New York, 2009)
26. A Raftery, M Newton, J Satagopan, P Krivitsky, in *Bayesian Statistics*. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity, vol. 8, (2007), pp. 1–45
27. D Lee, N Chia, A particle algorithm for sequential Bayesian parameter estimation and model selection. IEEE Trans. on Sign. Proc. **50**(2), 326–336 (2002)
28. H Mallick, N Yi, Bayesian methods for high dimensional linear models. J. Biom. Biostat. **1**(5), 326–336 (2013)
29. F Adjed, JF Giovannelli, A Giremus, N Dridi, P Szacherski, in *ser. Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2013)*. Variable selection for a mixed population applied in proteomics (Vancouver, 2013), pp. 1153–1157
30. C Robert, in *The Bayesian Choice. From decision-theoretic foundations to to computational implementation*. Springer Text in Statistics (Springer Verlag, New York, 2007)
31. G Saporta, *Probabilités, analyse de données et statistique*, Technip, Ed.Editions TECHNIP, (1990)
32. A Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distributions. Indian J Stat. **7**(4), 401-406 (1946)
33. L Song, A Smola, A Gretton, J Bedo, K Borgwardt, Feature selection via dependence maximization. J. Mach. Learn. Res. **13**(1), 1393–1434 (2012)
34. P Szacherski, JF Giovannelli, L Gerfault, P Mahé, JP Charrier, A Giremus, B Lacroix, P Grangeat, Classification of proteomic MS data as Bayesian solution of an inverse problem. IEEE Access. **2**, 1248–1262 (2014)
35. A Klich, C Mercier, L Gerfault, P Grangeat, C Beaulieu, E Degout-Charmette, T Fortin, P Mahé, JF Giovannelli, JP Charrier, A Giremus, D Maucort-Boulch, P Roy, Experimental design and statistical analysis for evaluation of quantification performance of two molecular profile reconstruction algorithms used in selected reaction monitoring-mass spectrometry. Service de Biostatistique, Hospices Civils and Laboratoire de Biométrie et Biologie Evolutive, Lyon, Technical Report (2016)
36. J Lemoine, T Fortin, A Salvador, A Jaffuel, JP Charrier, G Choquet-Kastylevsky, The current status of clinical proteomics and the use of MRM and MRM3 for biomarker validation. Pharmacogenomic Proteomic Metabolomic Appl. **12**(4), 333–345 (2012)